International Journal of Technology Assessment In Health Care, 6 (1990), 31–55.

A BAYESIAN METHOD FOR SYNTHESIZING EVIDENCE

The Confidence Profile Method

David M. Eddy
Vic Hasselblad
Ross Shachter
Duke University

Abstract

This article describes a collection of meta-analysis techniques based on Bayesian statistics for interpreting, adjusting, and combining evidence to estimate parameters and outcomes important to the assessment of health technologies. The result of an analysis by the Confidence Profile Method is a joint posterior probability distribution for the parameters of interest, from which marginal distributions for any particular parameter can be calculated. The method can be used to analyze problems involving a variety of types of outcomes, a variety of measures of effect, and a variety of experimental designs. This article presents the elements necessary for analysis, including prior distributions, likelihood functions, and specific models for experimental designs that include adjustment for biases.

An essential step in any technology assessment is the evaluation and synthesis of evidence to estimate parameters that describe the effect of the technology on important outcomes. These parameters can represent health outcomes,[1] intermediate outcomes,[2] economic outcomes, features of the population (e.g., prevalence rate of a disease, relative risk of a risk factor), or operating characteristics of the technology (e.g., sensitivity of a diagnostic test).

Occasionally there is a single experiment,[3] perfectly designed and conducted, that is directly applicable to the assessment problem. In such a case, it is reasonable to use the result of the experiment as a best estimate of the parameter and to use its confidence intervals to describe a range of uncertainty. More often, however, there are multiple studies, with different results, different designs and sizes, subject to a variety of biases, conducted in settings that differ slightly from one another and from the setting of interest. In addition, the evidence is often indirect, forcing the assessor to build models and piece together information on the population, the disease, and the technology. Finally, there are almost always gaps in the available evidence. The gaps can range from a complete absence of experiments about crucial parameters, to uncer-

The authors acknowledge the contributions of Robert Wolpert to this work and thank The John A. Hartford Foundation and the National Center for Health Services Research and Health Care Technology Assessment for their support.

tainty about the magnitude of a bias that affects the interpretation of an experiment. The synthesis of evidence of this type has been called meta-analysis.

This article describes a collection of meta-analysis techniques based on Bayesian methods for interpreting, adjusting, and combining evidence to estimate parameters and outcomes important to the assessment of health technologies. The techniques, which, taken together, are called the Confidence Profile Method, were introduced by David Eddy (8) and extended by Vic Hasselblad, Ross Shachter, Robert Wolpert, and Eddy.

FORMULATION OF THE ASSESSMENT PROBLEM

Application of the Confidence Profile Method begins with definitions of (a) the health problem (e.g., disease, patient population, indications), (b) the technology to be assessed (e.g., technique, dose, type of provider), (c) the alternative(s) with which it is to be compared, (d) the outcome(s) to be estimated, and (e) any features of the setting that could modify the effect of the technology on the outcomes. Collectively, these define the "circumstances of interest" and the parameters, θ, to be estimated in the assessment (the "parameters of interest").

USE OF EVIDENCE

The role of evidence is to help estimate the parameters of interest. When one is using experiments for this purpose, two problems must be addressed. First, because of random effects and sampling, the results of an experiment will not measure precisely the parameter we are trying to estimate. Second, the parameter actually being estimated by a particular experiment (called the "study parameter") might not match precisely any of the parameters of interest. This might occur because some circumstances of the study do not match the circumstances of interest (which threatens external validity) or because of the way the study was conducted or reported (internal validity). For a simple example of the first type, if we are interested in the effect of tamoxifen in a community setting, an experiment conducted in a research setting might be estimating a different parameter.

The consequences of these two characteristics of experiments is that the assessor must examine the circumstances of each experiment to determine the extent to which they match the circumstances of interest. When differences are found, either the study must be rejected, or its results must be adjusted to allow for the differences. Methods for performing these adjustments are an important feature of the Confidence Profile Method.

BAYES' FORMULA

Bayesian methods provide an attractive approach to the assessment of health technologies because they correspond to the way we think about assessment problems intuitively. At the start of any assessment, we have a prior understanding of the parameters of interest (e.g., the probability of a postoperative infection). We then examine each piece of evidence to learn the new information it provides about the parameters. On the basis of this new information, we revise our estimates of the parameters. We can repeat this process for each piece of evidence until, after all the evidence has been evaluated, we arrive at our final understanding of the parameters. This understanding incorporates our prior knowledge plus all the evidence that has been evaluated. Should

more evidence become available at a later time, that "final" understanding can become our prior understanding for a new round of analysis.

Bayesian methods provide a formal framework for accomplishing this series of steps. For example, suppose we are interested in a particular parameter (called θ). Our knowledge of the parameter of interest at any time will be described by a probability distribution. Our knowledge before evaluating the new evidence is described by a "prior distribution," which we will denote as $\pi(\theta)$ (π, for prior). The results of the experiments (which we will denote, in general, as X) contain information about θ. In the Bayesian approach, this information is captured in a "likelihood function," which gives the likelihood that the actual results of the experiment (X) would have occurred, for any particular hypothesis about θ. The likelihood function will be written as $L(X|\theta)$. After the evidence has been evaluated, our new knowledge of θ is described by a "posterior" distribution for θ, conditioned on X (after evaluating the evidence, X). This posterior distribution is written $\pi(\theta|X)$.

Bayes' formula can be used to calculate the posterior distribution as the product of the likelihood function and the prior distribution,

$$\pi(\theta|X) = kL(X|\theta)\pi(\theta) \tag{1}$$

where k is a normalizing constant:

$$k = 1/\int L(X|\theta)\pi(\theta)d\theta \, .$$

Additional pieces of evidence can be incorporated in the assessment by repeated applications of Bayes' formula. For example, suppose there is a second independent piece of evidence, with results X_2. After the first piece of evidence has been evaluated, but before evaluation of the second piece of evidence, our knowledge about θ is encoded in $\pi(\theta|X)$. This posterior distribution (derived after evaluation of the first experiment) can be used as the prior distribution for evaluation of the second piece of evidence, because it describes our knowledge before examining that piece of evidence. The information in the second experiment is captured in the likelihood function, $L(X_2|\theta)$. A revised posterior distribution for θ based on both pieces of evidence ($\pi(\theta|X,X_2)$) can be derived by a second application of Bayes' formula.

$$\pi(\theta|X,X_2) = k'L(X_2|\theta)\pi(\theta|X) \tag{2}$$

If we substitute for $\pi(\theta|X)$ from (1), we obtain

$$\pi(\theta|X,X_2) = k'L(X_2|\theta)L(X|\theta)\pi(\theta)^4 \tag{3}$$

EXAMPLE

Suppose the parameter of interest (θ) is the probability that women who receive modified radical mastectomies for stage II breast cancer can be discharged within 7 days postoperatively. Suppose that, before evaluation of any evidence, we profess to have no knowledge whatsoever of this probability. In such a case, we would represent our prior state of information with a distribution that carries no information about the value of the parameter. Such a distribution is the beta distribution with parameters $\alpha = ½$, $\beta = ½$: $(\theta^{-0.5}(1-\theta)^{-0.5})/\pi$.

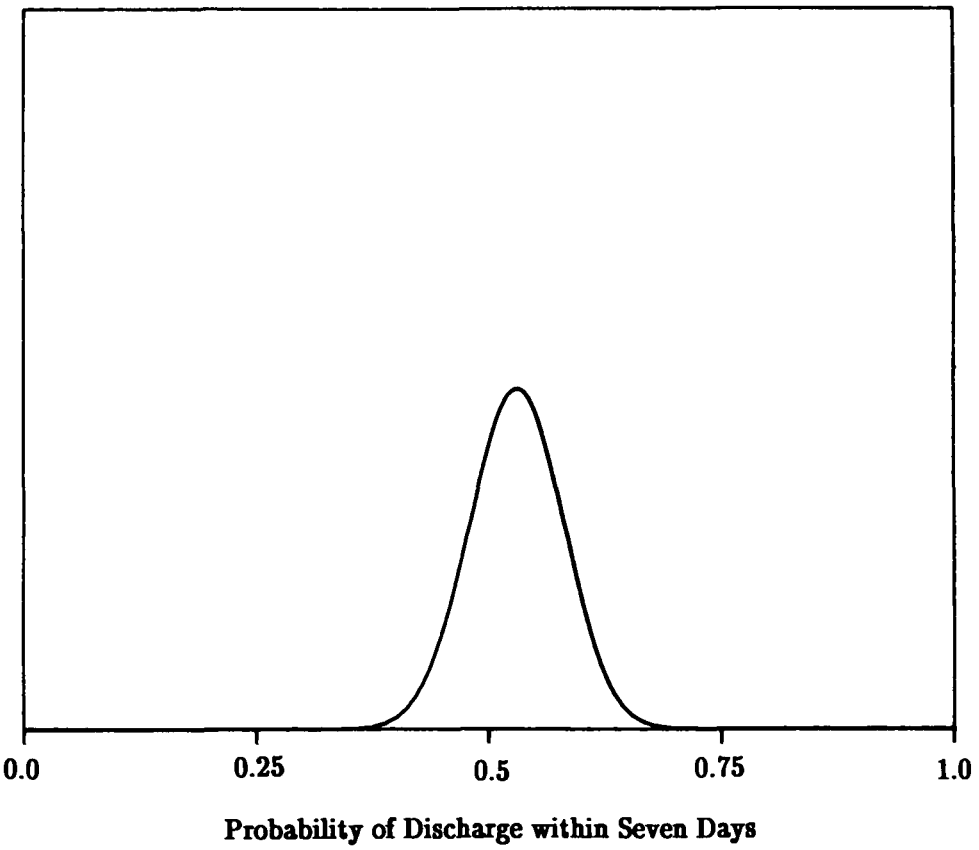

Figure 1. Graph of likelihood function for probability of discharge within 7 days, based on evidence from experiment.

Now suppose there is an experiment in which 100 women were treated, and 53 were observed to be discharged within 7 days. The likelihood function calculates the likelihood that, for any particular value of θ, 53 of 100 women actually would be discharged by 7 days. The likelihood function for this type of experiment is based on the binomial distribution and is given by

$$L(53 \text{ of } 100|\theta) \propto \theta^{53}(1 - \theta)^{47}$$

The proportional sign is used because likelihood functions are determined only up to an arbitrary multiplicative constant. This function is illustrated in Figure 1.

Equation (**1**) can be used to calculate a posterior distribution for θ that combines the prior knowledge ($\pi(\theta)$) and the information in the experiment ($L(X|\theta)$). The result is

$$\pi(\theta|53 \text{ of } 100) = k''\theta^{53}(1 - \theta)^{47}\,\theta^{-0.5}\,(1 - \theta)^{-0.5} \qquad (4)$$

where k'' is the normalizing constant, $\Gamma(101)/\Gamma(53.5)\Gamma(47.5)$, and Γ is the standard gamma function. This is illustrated in Figure 2.

Now suppose there is a second, larger experiment in which 212 women were fol-

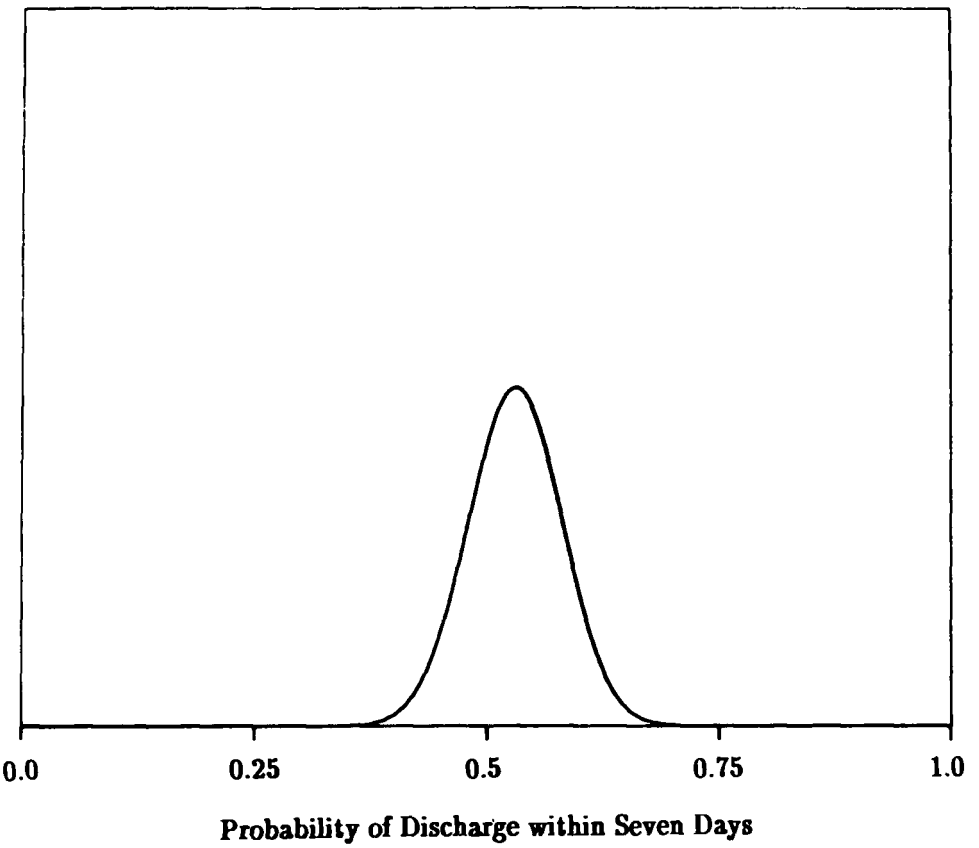

Figure 2. Graph of posterior probability distribution for probability of discharge within 7 days, based on evidence from experiment and noninformative prior distribution.

lowed, and 130 were observed to be discharged by the seventh day. The likelihood function for this experiment is again derived from the binomial distribution

$$L(130 \text{ of } 212|\theta) \propto \theta^{130}(1 - \theta)^{82} \tag{5}$$

This is illustrated in Figure 3.

A second application of Bayes's formula using the distribution for θ obtained after evaluation of the first piece of evidence (**4**), and the likelihood function for the second piece of evidence (**5**), yields a posterior distribution for θ that takes into account both pieces of evidence.

$$\pi(\theta|53 \text{ of } 100 \text{ and } 130 \text{ of } 220) =$$
$$k'''\theta^{53}(1 - \theta)^{47}\theta^{130}\,(1 - \theta)^{82}\,\theta^{-0.5}(1 - \theta)^{-0.5}$$

where k''' is the normalizing constant. This is illustrated in Figure 4.

If there were additional pieces of information, they could be incorporated in the assessment by repeated applications of the formula. After all of the evidence has been

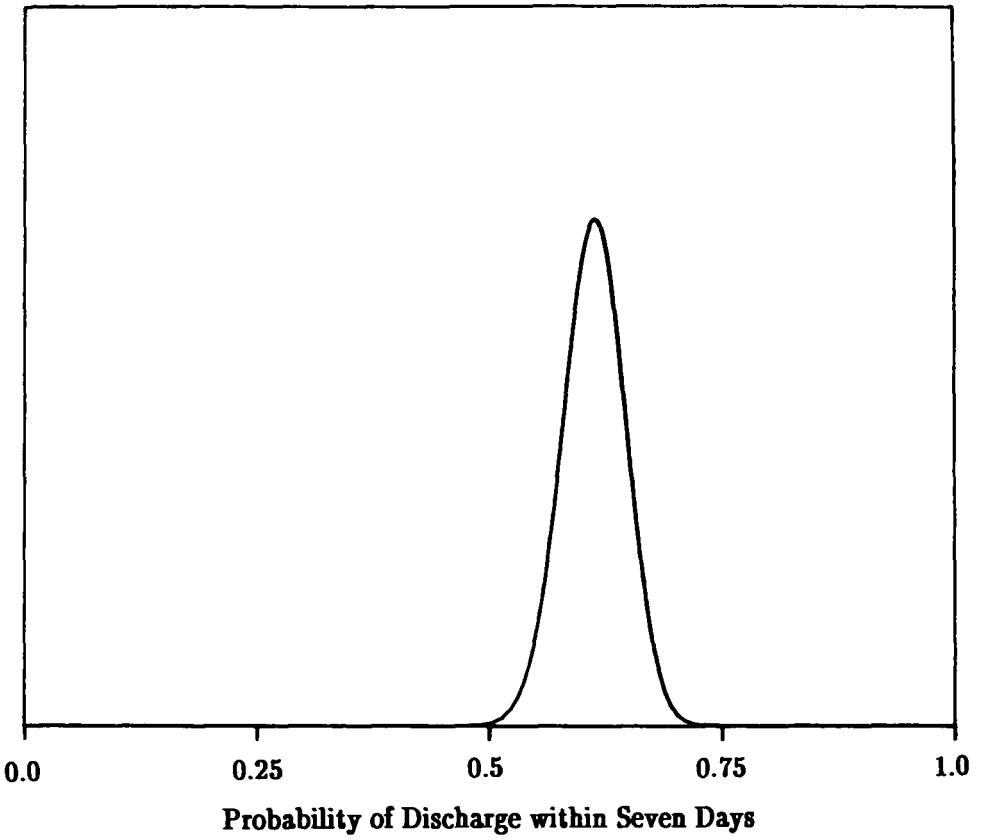

Figure 3. Graph of likelihood function for probability of discharge within 7 days, based on evidence from second experiment.

considered, the final posterior distribution can then be used in subsequent steps of the technology assessment (e.g., a mathematical model).

APPLYING BAYESIAN METHODS

The Confidence Profile Method involves building a model that relates parameters and evidence. There are three elements: basic parameters, functional parameters, and likelihood functions. A parameter is basic if it is not defined as a function of any other parameter in the model. Basic parameters can be thought of as parameters handed to us by Mother Nature. An example is the "true" rate of an outcome in the control group of a particular experiment (θ_0). In a Bayesian analysis, all basic parameters have prior distributions. A functional parameter is a parameter that is defined as a function of other parameters. For example, the difference (θ_d) between the outcome rate in the control group (θ_0) and treated group (θ_1) of a trial is a functional parameter defined by $\theta_d = \theta_1 - \theta_0$. The relationships between evidence and parameters (either basic or functional) are defined by likelihood functions. Together, the parameters and the functions that relate them form a system of equations, or a model.

After the assessment problem has been formulated and the parameters have been defined, the construction of a model involves several steps. First, assessors must either be able to choose a noninformative prior or specify a prior distribution to represent

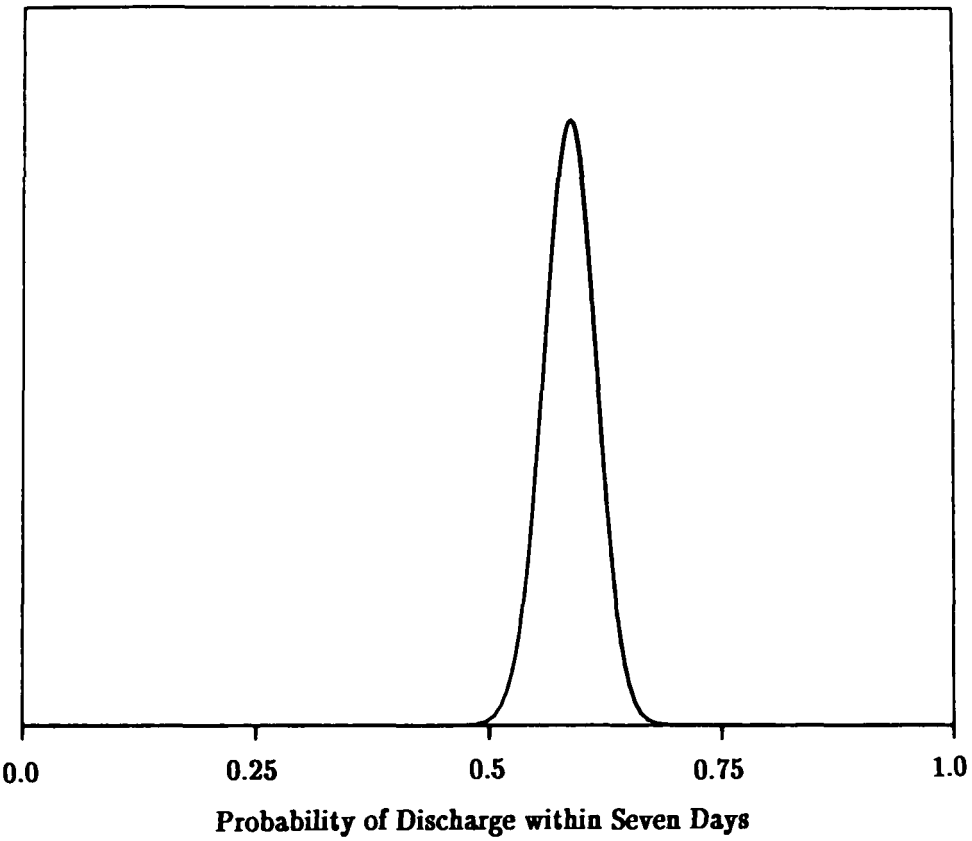

Figure 4. Graph of posterior probability distribution for probability of discharge within 7 days, based on evidence from both experiments and noninformative prior distribution.

their knowledge about the basic parameters, prior to evaluation of any evidence. Second, for any type of experimental design, there must be a likelihood function that captures the information in the experiment about any of the parameters of interest. These likelihood functions must be able to accommodate different types of outcomes (e.g., dichotomous, continuous) and different measures of the effect of a technology on outcomes (e.g., differences, ratios, odds ratios). Third, for each study the assessor must (a) be able to adjust for any biases to internal or external validity, (b) be convinced that there are no biases, or (c) omit the experiment from the assessment.

PRIOR DISTRIBUTIONS

If a Bayesian analysis for a parameter of interest has been conducted previously, the resulting posterior distribution can be used as a prior distribution for the parameter for the next application of the model. If no evidence has yet been analyzed, the assessor has two choices. One is to use subjective judgments. When this is done, great care must be taken to ensure that these judgments do not incorporate any of the information that will subsequently be included in the assessment. Obviously, great care must also be taken to ensure that the judgments are justified by any preexisting evidence. Uncritical use of subjective judgments can lead to errors and abuse.

A safer and more conservative approach is to use a "noninformative prior" that

contains no information about the parameter of interest. Use of a noninformative prior is equivalent to stating that, before any evidence has been evaluated, the assessor favors no possible values of the parameters over others. James O. Berger (2) has described several methods for determining noninformative priors. In general, we have used Jeffreys' priors where appropriate. For parameters defined on the entire real line, the location-invariant noninformative prior is

$$\pi(\theta) = 1, \qquad \text{for } \theta \in (-\infty, \infty)$$

This distribution is "improper" because its integral does not equal 1. For parameters defined on an interval $\theta \in (0, \infty)$, the scale-invariant prior (also improper) is

$$\pi(\theta) = 1/\theta, \qquad \text{for } \theta \in (0, \infty)$$

which is also improper. For probabilities defined on the interval (0,1), the method of Harold Jeffreys (9) gives a beta distribution, with parameters $\alpha = \frac{1}{2}$, $\beta = \frac{1}{2}$. For the multinomial model, the analogous prior for the θ is a Dirichlet, with parameters $\frac{1}{2}, \frac{1}{2}, \ldots, \frac{1}{2}$.

The need to assign prior distributions for parameters is, for many people, a troublesome feature of Bayesian methods. The fear is that the choice of an inappropriate prior distribution might distort the results of the assessment (the posterior distribution). Fortunately, the impact of the prior distribution on the results of an assessment is inversely proportional to the amount of information available through experimental evidence. Furthermore, the impact of the prior distribution rapidly diminishes in the presence of relatively small amounts of evidence. For example, in the presence of just a single experiment worthy of analysis by classical statistics, the posterior distribution is quite insensitive to the choice of any prior that could reasonably be called "noninformative." However, when there is extremely little empirical evidence, the choice of a prior can affect the results. When an assessor suspects the choice of a prior distribution might influence the results, a prudent approach is to explore the sensitivity of the results to various choices. If the results are sensitive and if the assessor is not confident about which prior distribution is appropriate, the consequences of different options can be described, or in the extreme and rare case, the assessment can be aborted.

LIKELIHOOD FUNCTIONS

Likelihood functions relate parameters to evidence. The Confidence Profile Method uses likelihood functions for a variety of types of outcomes, effect measures, and experimental designs (see Table 1).

Outcome Measures

There are four main types of outcomes: dichotomous, categorical, counts, and continuous. A dichotomous outcome is an event that either occurs or does not. Examples are survival 5 years after a surgery for thyroid cancer (the patient either survives or not), return to work within a year after a myocardial infarction (MI) (the patient returns to work or not), or the occurrence of a postoperative infection (the patient has a postoperative infection or not). Categorical outcomes are events that can take one of several values (a finite number). Examples are four stages of a cancer, "mild," "moderate," or "severe" pain, and four degrees of disability following a stroke. Counts, as the name implies, count the occurrence of repeatable events. Examples are the number

Table 1. Likelihood Functions for Various Types of Experimental Designs, Outcomes, and Effect Measures

	Outcomes			
Designs	Dichotomous	Categorical	Count	Continuous
One-arm	Rate	Rate Score	Mean count	Mean score Median score
Two-arm	Difference Odds ratio % Difference	Difference Ratio	Difference Ratio	Difference Ratio
n-arm	Coefficients of logistic regression equation, β_i	Coefficients of linear regression equation, β_i	Coefficients of linear regression equation, β_i	Coefficients of linear regression equation, β_i
2 × 2 CCS	Odds ratio	NA	NA	NA
2 × *n* CCS	Coefficients of logistic regression equation, β_i	NA	NA	NA
Matched CCS	Odds ratio	NA	NA	NA
Cross-sectional	Coefficients of logistic regression equation, β_i	Coefficients of linear regression equation, β_i	Coefficients of linear regression equation, β_i	Coefficients of linear regression equation, β_i

NA: not available.

of requests a patient makes for extra pain relief following an operation, the number of follow-up visits per year for a glaucoma treatment, and the number of telephone calls required to get a clinic appointment. Examples of continuous outcomes are changes in life expectancy, the cost of a procedure, or duration of hospitalization.

Effect Measures

The purpose of a technology assessment is to estimate the effect of a technology on important outcomes. This requires a comparison of outcomes that occur with the technology versus the outcomes that occur with the specified alternative. Depending on the type of outcome, there can be different measures of the effect of the technology, compared with the alternative.

Dichotomous. Let θ_1 be the probability of the outcome in the group receiving the technology (the "treated" group) and θ_0 be the probability of the outcome in the group receiving the alternative (the "control" group). There are four main measures of effect, which we will denote as ε:

- absolute difference in probabilities, $\varepsilon = \theta_1 - \theta_0$;
- relative risk, $\varepsilon = \theta_1/\theta_0$;
- odds ratio, $\varepsilon = (\theta_1/(1 - \theta_1))/\theta_0/(1 - \theta_0)$; and
- percent change, $\varepsilon = 100(\theta_1 - \theta_0)/\theta_0$.

Categorical. For categorical outcomes, let the probability of the i^{th} category be θ_i. A measure of the effect of changing the probabilities a patient will be in particular categories can be obtained by assigning a value to each category (denoted as α_i), and calculating a score as $S = \Sigma\alpha_i\theta_i$. The measure of effect is the difference in scores with and without the technology

$$\varepsilon = \sum_i \alpha_{1i}\theta_{1i} - \sum_i \alpha_{0i}\theta_{0i}$$

where the additional subscripts $_1$ and $_0$ designate the technology and the control.

For example, the "values" of cancer stages might be the 5-year survival rates of patients diagnosed in that stage. The score will then be the expected 5-year survival. Alternatively, the value for stages of a cancer could be the cost of initial treatment for each stage, with the score being the expected cost. If assessors are interested in how screening will change the probability a patient is diagnosed in stage I or II, they can assign values of 1, 1, 0, and 0, to stages I, II, III, and IV, respectively, which will cause the score to be the probability of a stage I or II cancer.

Counts. For counts, the most useful measures of effect are the difference in counts with and without the technology ($\varepsilon = c_1 - c_0$) or the ratio of counts ($\varepsilon = c_1/c_0$).

Continuous. Common measures of the effect of a technology on continuous outcomes are the absolute difference in means ($\varepsilon = \mu_1 - \mu_0$), the ratio of means ($\varepsilon = \mu_1/\mu_0$), or the difference in medians ($\varepsilon = m_1 - m_0$). The "effect size" can also be calculated as $\varepsilon = (\mu_1 - \mu_0)/\sigma$, where σ is the standard deviation of the control group, or a "combined" standard deviation.

Experimental Designs

For the purposes of describing likelihood functions, it is convenient to classify experimental designs into seven groups: prospective one-arm, prospective two-arm, prospective multi-arm, 2×2 case control, $n \times 2$ case control, and cross-sectional studies. A prospective design involves identifying individuals exposed to the technology of interest (or an alternative technology), following the individuals over time, and collecting data on outcomes as they occur. The number of arms is determined by the number of groups exposed to different levels or types of the technology.

Case-control designs involve identifying individuals who have either had the outcome of interest ("cases") or not ("noncases"). By definition, such a study involves only dichotomous outcomes (cases or noncases). After the cases and noncases have been identified, the investigators look back to determine which individuals had been exposed to the technology. In a 2×2 case-control study, exposure is dichotomous; an individual is determined to be either "exposed" or "not exposed." In a $2 \times n$ case-control study, the individuals can be exposed to different levels of intensity of the technology (e.g., different doses of a drug, different frequencies of screening, or different amounts of tobacco).

Cross-sectional designs take a snapshot of a population at a particular time to determine which individuals are (or were) exposed to the technology, and which have had the outcomes of interest. For each individual in the population, the investigator records the outcome of interest, exposure to the technology, and additional factors relating to the outcomes of interest (e.g., education, smoking history). Patients are not followed through time; the study only describes the correlation between various factors and outcomes at one particular time.

All these designs can concurrently collect data on other factors that might alter

the occurrence of the outcomes, exposure to the technology, or the effect of the technology on outcomes (covariates).

Prospective One-Arm

In a one-arm study, only a single group of patients is followed. The best example is a clinical series, although data bases and surveillance studies can be used for this purpose.

Dichotomous. The likelihood function for a one-arm prospective study with dichotomous outcomes is based on the binomial distribution. Let θ be the true rate of the event of interest, let s (for "success") be the number of occurrences of that event over a specified period of time, and let f (for "failure") be the number of nonoccurrences. The likelihood function is

$$L(s,f|\theta) \propto \theta^s(1 - \theta)^f.$$

The binomial model assumes the probability of a "success" (θ) is the same for each individual, and (conditional on θ) the outcome for any particular individual is independent of the actual outcome of any other individual.

Categorical. For categorical outcomes, let θ_i be the true probability the event of interest will occur in category i, and let s_i be the number of actual occurrences in category i. The likelihood function $L(s_i|\theta_i)$ is derived from the multinomial model, which assumes that the outcomes for each individual depend only on the underlying probabilities for each category (θ_i), and are otherwise independent of the actual outcomes of other individuals.

$$L(s_i|\theta_i) \propto \prod_i \theta_i^{s_i}$$

Counts. For counts, let λ be the expected number of counts over a specified period of time (the parameter of interest). Let c_i be the observed number of counts for the i^{th} unit. (A unit can be an individual, a county, a hospital, or any other entity to which the repeatable events can occur.) One possible likelihood function is derived from the Poisson model, which assumes independent interarrival times.

$$L(c_i|\lambda) \propto \frac{e^{-n\lambda}\lambda^{\Sigma c_i}}{\Pi c_i!}$$

where n is the total number of units. More complicated models involving negative binomial distributions are possible.

Continuous. The likelihood function for a one-arm prospective experiment involving continuous outcomes is based on the assumption that the underlying distribution for the data is normal or can be transformed into a normal distribution (e.g., log-normal), and that the individual observations (x_i) are independent. The parameters of interest are the mean (μ) and standard deviation (σ) of the normal distribution. If the original data were transformed to achieve normality (e.g., by taking the natural log), a posterior distribution for the parameter of interest in the untransformed space can be obtained by an appropriate retransformation (e.g., by taking the antilogarithm).

The likelihood function for the data (x_i) in terms of the mean and standard deviation is

$$L(X|\mu,\sigma) \propto \frac{1}{(2\pi\sigma)^{n/2}} e^{-\sum_i (x_i - \mu)^2/2\sigma^2}$$

With the use of appropriate noninformative prior distributions for μ and σ, the marginal posterior distribution for μ can be shown to be a student t distribution with parameters $\Sigma x_i/n(=\bar{x})$, $\Sigma(x_i - \bar{x})^2/(n-1)$, and $n-1$. A marginal posterior distribution for σ can also be derived.

Derivation of a likelihood function for a posterior distribution for the median of a continuously distributed parameter requires nonparametric methods. An approximate solution can be obtained using an approach outlined by Jeffreys (9).

Prospective Two-Arm

In a prospective two-arm experiment, two groups of individuals are followed, one exposed to the technology (sometimes called the "treated" group), the other exposed to a specified alternative (the "control" group). Examples are randomized controlled trials (RCTs), nonrandomized controlled trials (NRCTs), and observational studies. Information from data bases can sometimes be analyzed as a prospective two-arm study. The main differences between these designs is their vulnerability to biases to internal and external validity. Specific biases and methods of adjustment will be discussed separately.

Dichotomous. In the dichotomous case, the parameters of interest are θ_1 and θ_0, the true rates of the outcome in the treated and control groups, respectively. The results of the study are the number of successes and failures in the two groups (s_1,f_1,s_0,f_0). The likelihood function is again based on the binomial distribution

$$L(s_0,f_0,s_1,f_1|\theta_0,\theta_1) \propto \theta_0^{s_0}(1-\theta_0)^{f_0}\theta_1^{f_1}(1-\theta_1)^{f_1} \quad (6)$$

Likelihood functions for various effect measures (e.g., relative risk, odds ratio, difference in rates) can be derived by (a) defining a general effect measure as a function of θ_0 and θ_1, $\varepsilon = H(\theta_0,\theta_1)$; (b) solving this for θ_1 in terms of θ_0 and ε, $\theta_1 = H^{-1}(\theta_0,\varepsilon)$; (c) substituting for θ_1 in (6); (d) specifying a (possibly noninformative) prior distribution for $\theta_0(g(\theta_0))$; and (e) integrating over θ_0 (1;3). Thus

$$L(s_0,f_0,s_1,f_1|\varepsilon) \propto$$
$$\int_0^1 \theta_0^{s_0}(1-\theta_0)^{f_0}(H^{-1}(\varepsilon,\theta_0))^{s_1}(1-H^{-1}(\varepsilon,\theta_0))^{f_1}g(\theta_0)d\theta_0$$

Categorical. For categorical outcomes, the parameters of interest are θ_{ji}, where j designates absence ($j=0$) or presence ($j=1$) of the technology, and i designates the i^{th} category. The results of the experiment are summarized in the number of cases observed in each category, for the two groups, s_{ji}. The likelihood function is given by

$$L(s_{ji}|\theta_{ji}) \propto \prod_j \prod_i \theta_{ji}^{s_{ji}}$$

If a Dirichlet distribution with parameters α_i is used as a prior distribution, the posterior distribution for θ_i is Dirichlet with parameters $(\alpha_i + s_i)$.

More complicated calculations are required to derive likelihood functions or posterior distributions for the scores of the two groups, which in the deterministic

case is given by $\varepsilon_j = \sum_i \alpha_{ji}\theta_{ji}$, $j = 0, 1$. However, the means and variances are easily calculated. For the j^{th} group,

$$\text{mean} \equiv M_j = \sum_i c_{ji}(\alpha_{ji} + S_{ji})/T_j$$

$$\text{var} = \sum_i c_{ji}^2(\alpha_{ji} + x_{ji})/T_j(T_j + 1) - M_j^2(T_j + 1)$$

where $T_j = \Sigma\alpha_{ji}X_{ji}$. A distribution for the overall effect of the technology, defined as the difference in scores ($\varepsilon = \varepsilon_1 - \varepsilon_0$) is calculated by computing the convolution of the distributions for ε_1 and ε_0.

Counts. The likelihood function for two-arm prospective studies in which the outcomes are counts is similar to the formula for the one-arm case. Specifically, let λ_j, $j = 0, 1$ be the true expected number of counts per unit (e.g., time interval) in the control and treated groups, and let c_{ji} be the number of counts for the i^{th} observation (e.g., individual) in the two groups ($j = 0, 1$). Then the joint likelihood function for λ is given by

$$L(c|\lambda) \propto \prod_j \frac{e^{-n\lambda_j}\lambda_j^{\sum_i c_{ji}}}{\prod_i c_{ji}!} \tag{7}$$

Likelihood functions and posterior distributions for various effect measures, such as the difference in λs and the ratio of λs, can be calculated by specifying an effect function of the form $\varepsilon = H(\lambda_0,\lambda_1)$, solving for λ_1 in terms of λ_0 and ε, substituting for λ_1 in (7), specifying a (possibly noninformative) prior distribution for λ_0, and integrating over λ_0.

Continuous. Likelihood functions for two-arm prospective studies with continuous outcomes are natural extensions of the one-arm case. A general formula for n-arms will be presented below.

Prospective Multi-Arm

Multi-arm prospective studies can be separated into two types. In one, there is no apparent model relating the effects of the technologies received by each group. For example, a multi-arm controlled trial of various nonsteroidal anti-inflammatory agents (NSAIDs) might have three groups, each getting three different NSAIDs, as well as a placebo group. The analysis of this type of multi-arm study consists of analyzing various pairs of groups, such as comparing NSAID #1 versus the placebo, NSAID #1 versus NSAID #2, and so forth. The methods already described for two-arm studies can be used to perform these assessments.

The other type of multi-arm study arises when the assessor is willing to postulate an underlying model that relates the effects of the technologies in the various groups. For example, a multi-arm design might be used to study several different doses of a particular NSAID, different frequencies of screening, different amounts of fat in the diet, or different amounts of a carcinogen. This type of study is especially powerful because, if a model can be specified to relate the outcome to the "intensity" of the technology in each arm, the results of each arm will provide information about the effects of the other arms, as well as about other intensity levels not specifically included in the study. For example, a four-arm drug study involving a placebo, 50 mg, 100 mg, and 200 mg, carries information about the effectiveness of 150 mg. The likeli-

hood functions for this type of multi-arm study depend on the models that relate the different intensities of the technology.

Dichotomous. For multi-arm studies likelihood functions have been derived using a multiple logistic regression model. Let the rate of the event for the i^{th} arm (e.g., dose i) be given by $\theta_i = \text{logit}(\beta_0 + \beta_1 x_i)$, where $\text{logit}(z) = 1/(1 + e^{-z})$, β_0 is the log of the odds for the control group (no technology), β_1 is the log of the odds ratio for the technology with intensity "1" (the reference intensity), and x_i is the dose in the i^{th} group. The model can be expanded to incorporate additional factors, such as age, sex, race, or risk factors. The general form is $\theta_i = \text{logit}\ (\Sigma\beta_j x_{ij})$ with j indexing the independent variables. The results of the study are summarized by the number of successes and failures in each arm (s_i and f_i).

The likelihood function for the s_i and f_i conditional on the θ_i is

$$L(s_i f_i|\theta_i) = \prod_i \theta_i^{s_i}(1 - \theta_i)^{f_i}$$

Because the model is completely specified by the coefficients β_j, it is desirable to derive posterior distributions for the β_j. One approach is to substitute for the θ_i in terms of the β_j and derive a likelihood function for the results of the experiment in terms of the β_j. Another is to exploit the property that, near the values of β_j that maximize the likelihood function, the likelihood function is approximately normal, with a mean given by the maximum likelihood estimates of the β_j and a variance that can be calculated from the second partial derivatives of the likelihood function (5).

Categorical. Multi-arm studies with categorical outcomes are very complicated and, fortunately, extremely rare. We have not derived likelihood functions for this type of design.

Counts. Likelihood functions have been published for multi-arm prospective studies involving count outcomes. An example of this type of design is the Ames test, in which the investigator counts the number of mutant colonies that occur when bacteria are exposed to various doses of potential carcinogens. A likelihood function has been derived based on the Poisson model, with the use of a regression model for dose (12).

Continuous. Multi-arm studies with continuous outcomes can be analyzed using a regression model

$$y_i = \sum_j \beta_j x_{ij} + \xi_i$$

where y_i is the value of the dependent variable (e.g., blood pressure level) for the i^{th} case (e.g., individual), x_{ij} are the values of the j^{th} factor (e.g., treatment dose, age, sex, smoking history) for the i^{th} individual, β_j are the continuous coefficients for the j^{th} factor ($j = 1$ to k), and ξ_i is an error term for the i^{th} observation. The ξ_i are assumed to have independent normal distributions, all with mean 0, variance σ^2.

The likelihood function is most easily written in matrix form,

$$L(Y|\bar{\beta},\tau) \propto \tau^{n/2} e - \tau(Y - X\bar{\beta})^T(Y - X\bar{\beta})/2$$

where $\tau = 1/\sigma^2$, and Y, X, and $\bar{\beta}$ are the vectors of ys, xs, and βs.

DeGroot (5) has shown that under a reasonable choice of noninformative priors

for τ and $\bar{\beta}$, the posterior distribution for the β_j, conditional on the observed results of the study (y_i), is a multivariate student t distribution. The location parameter is

$$m = (X^TX)^{-1}(X^TY)$$

This corresponds to the classical least squares estimates of the β. The precision matrix[5] is $(n - k)/[(Y - X^Tm)^T(Y - X^Tm)(X^TX)^{-1}$, and the degrees of freedom of the t distribution is $n - k$.

From this, a marginal posterior distribution can be obtained for the regression coefficient. For the simple regression model, joint and marginal posterior distributions can be obtained for the slope and intercept.

2 × 2 Case-Control Studies

The parameter of interest in a 2 × 2 case-control study is the odds ratio, ε or ε_r.

$$\varepsilon_r = (\theta_1/(1 - \theta_1))/\theta_0/(1 - \theta_0) \tag{8}$$

where θ_1 is the proportion of cases exposed to the technology and θ_0 is the proportion of noncases exposed. The results of a case-control study are summarized in a 2 × 2 table as illustrated in Table 2.

The likelihood function for the x_{ij}, conditional on θ_1 and θ_0, is based on the binomial distribution.

$$L(X_{ij}|\theta_0\theta_1) = \theta_1^{X_{11}}\theta_0^{X_{12}}(1 - \theta_1)^{X_{21}}(1 - \theta_0)^{X_{22}} \tag{9}$$

A likelihood function in terms of the odds ratio can be obtained by solving (8) for θ_1 in terms of θ_0 and ε_r, substituting in (9), and integrating over θ_0.

Matched Case-Control Study

In a matched case-control study, each case is matched with a specified number of controls according to specified factors such as age, sex, race, and other possible confounding variables. The results are summarized as illustrated in Table 3: x_{1j} is the number of exposed cases for which j of the matched controls were exposed, x_{2j} is the number of nonexposed cases for which j of the matched controls were exposed, and n is the number of controls matched to each case. The likelihood function for the odds ratio (4) is

$$L(X_{ij}|\varepsilon_r) \propto \prod_{i=1}^{n} [i\varepsilon_r/(i\varepsilon_r + n - i + 1)]^{X_{1i}} [(n - i + 1)/(i\varepsilon_r + n - 1 + 1)]^{X_{0i}}.$$

2 × *n* Case-Control Study

This design allows for different levels of intensity of the technology (e.g., dose of a drug, cigarette packs per day). For each case and control, the investigator records the levels of intensity of the exposure, as illustrated in Table 4: s_{i1} is the number of cases exposed to intensity level x_i, and s_{i2} is the number of noncases exposed to intensity level x_i.

The parameters of interest, ϕ_i, are the proportions of subjects exposed to technology level i (denoted as x_i) who are cases. The parameters ϕ_i are modeled as logistic functions with coefficients β_0 and β_1. That is,

$$\phi_i = \text{logit}(\beta_0 + \beta_1 x_i)$$

Table 2. Results of a 2 × 2 Case-Control Study

Status of Cases	Cases	Noncases
Exposed	x_{11}	x_{12}
Not exposed	x_{21}	x_{22}

Table 3. Results of a Matched Case-Control Study

	Number of Exposed Controls				
Status of Cases	0	1	2	. . .	*n*
Exposed	x_{10}	x_{11}	x_{12}	. . .	x_{1n}
Not exposed	x_{20}	x_{21}	x_{22}	. . .	x_{2n}

Table 4. Results of a 2 × *n* Case-Control Study

Level of Exposure to the Technology	Case	Not Case
X_0	s_{01}	s_{02}
X_1	s_{11}	s_{12}
X_2	s_{21}	s_{22}
X_n	s_{n1}	s_{n2}

In this formulation, β_1 is the log of the odds ratio of being a case for someone exposed to the reference intensity level 1 (compared with no exposure to the technology). The likelihood function is given by

$$L(s|\phi) \propto \prod_j \phi_j^{s_{j1}}(1 - \phi_j)^{s_{j2}}$$

which can be solved to obtain a likelihood function in terms of β_0 and β_1.

Cross-Sectional Studies

The likelihood functions for this group of designs depend on the type of the outcome (dichotomous, categorical, continuous), which determines the appropriate model relating the dependent and independent variables. For dichotomous outcomes, the analysis is based on the multiple logistic regression model, as described for the multiple-arm prospective study with dichotomous outcomes. For categorical outcomes, a sequence of multiple logistic regression models can be used. Cross-sectional designs with counts or continuous outcomes can be analyzed using a linear regression model, as described under multi-arm prospective studies. Thus, the analysis of cross-sectional studies is analogous to the multi-arm prospective studies, with appropriate modifications in the definitions of the dependent and independent variables.

ADJUSTMENTS FOR BIASES

Biases threaten virtually every study. This forces the assessor to make a series of judgments. First, the assessor must decide whether to (a) accept the study at face value, implying that the biases are so small that their presence will not materially alter the results; (b) reject the study altogether, implying that the biases are so large that the

study is totally without value; or (c) adjust the results of the study to take into account the potential biases. If the assessor chooses the latter course, there is a second set of judgments involving the magnitudes of the biases, and the appropriate models for performing adjustments.

The identification of potential biases begins with the formulation of the assessment problem – the definitions of the health problem, technology, comparisons, outcomes, and setting. These factors determine the circumstances of interest, and the parameters of interest, θ. Then, for each experiment, the assessor must evaluate the same factors. For example, what was the health problem under investigation, what was the technology, and so forth? The answers to these questions define, for each study, the circumstances of the study, and the study parameter.

If the circumstances of the study differ from the circumstances of interest in any way that might affect outcomes, a decision must be made about acceptance, rejection, or adjustment of the study. If the latter is chosen, the assessor must understand that the study parameter (call this θ') is different from the parameter of interest (θ), and (1) or (2) cannot be applied without modification. Specifically, the likelihood functions derived in such circumstances will apply to the study parameter. To proceed, it is necessary to define the relationships between the study parameter and the parameter of interest. The functions that define these relationships can then be used to derive adjusted likelihood functions, written in terms of the study parameter, or to expand the model to include the study parameter as a function of the parameter of interest.

The Confidence Profile Method includes specific models for the most common biases to internal and external validity.

Biases to Internal Validity

There are four main types of biases to internal validity: error in measurement of outcomes, protocol departures, patient-selection biases, and misreporting of results.

Error in Measurement of Outcomes. In any experiment, there is a possibility that the method or instrument used to measure outcomes might be erroneous. For examples, the cause of death listed on death certificates, patient interviews to determine alcohol abstinence, urine tests for drugs, blood pressure readings, and IQ tests can all be inaccurate. In retrospective studies, this bias appears as an incorrect classification of "cases" versus "not cases."

To model this bias for experiments involving dichotomous outcomes, for each arm let α be the probability that a true success will be incorrectly labeled a failure, and let β be the probability that a true failure will be incorrectly labeled a success in that arm. For that arm, the relationship between the study parameter (θ') and the parameter of interest (θ) is

$$\theta' = (1 - \alpha)\theta + \beta(1 - \theta)$$

The α and β can be different for different arms.

A measurement bias involving continuous outcomes is equivalent to calibration error. Let α and β be arbitrary real numbers. Two possible models relating θ' and θ are

$$\theta' = \alpha + \beta\theta$$

or

$$\theta' = \alpha\theta^{\beta}$$

Protocol Departures. Protocol departures in prospective studies occur either when individuals intended to get the technology do not (receiving instead a different technology or no technology), or individuals intended not to get the technology, do. The latter is called "contamination" of the control group, and the former "dilution" of the treated group. In retrospective studies, protocol departure takes the form of misspecification of exposure to the technology. For example, when reviewing the records of cases or controls, researchers might erroneously indicate that a patient received the technology when in fact he or she did not, or vice versa.

To model this bias, designate α as the probability that a person allocated to the treated group (or labeled as "treated") does not in fact receive the treatment, and, as before, let θ_1 and θ_0 be the probability of the event of interest in people who actually receive the technology, or do not (controls), respectively. The probability of an event in the group offered treatment (including the "dilutants") is

$$\theta_1' = (1 - \alpha)\theta_1 + \alpha\theta_0$$

To model contamination, let β be the probability a person in the control group receives the treatment. Then

$$\theta_0' = \beta\theta_1 + (1 - \beta)\theta_0$$

Notice that both θ_1 and θ_0 might themselves require adjustment if the people who dilute or contaminate are subject to a selection bias or to a technology bias.

Patient-Selection Bias. Patient-selection bias occurs in two-arm or multi-arm studies when the individuals in the two groups have inherently different probabilities of the outcomes of interest, irrespective of their exposure to the technology(s). For example, a two-arm study comparing cancer incidence rates in two factories, one of which uses a suspected chemical carcinogen, might be biased if the employees of one factory smoke more or are older than the employees of the other factory. Either smoking or age could affect cancer incidence rates independently of exposure to the carcinogen (or could act synergistically with the carcinogen). For another example, patients offered an experimental treatment might have more severe disease than those not offered the treatment.

A general model for the relationship between the study parameter (θ') and the parameter of interest (θ) can be defined as

$$\theta' = \alpha + \beta\theta'$$

where α and β are any real numbers.

More specific models can be developed to help estimate the parameters α and β. For example, if the control and treated groups differ with respect to several factors (e.g., sex, age, education, tobacco use), a logistic regression model can be developed that enables the assessor to estimate an adjustment to the odds ratio, from estimates of the odds ratios assessed with each factor (assuming independence between factors). (These estimates correspond to the βs of the logistic regression model.) The latter can be estimated empirically or, if appropriate, subjectively.

Loss to Follow-Up. A frequent problem in longitudinal studies is that patients in the treated or control groups can be lost to follow-up. If those lost to follow-up

have the same risk of the outcome as patients who were not lost to follow-up, then no adjustment is necessary (other than the smaller sample size). On the other hand, if the patients lost to follow-up are not believed to have the same risk of the outcome as patients who were not lost, an adjustment is necessary. This can be accomplished by specifying an odds ratio that describes the risk in those lost to follow-up, compared with those not lost to follow-up. If desired, this odds ratio can be estimated with a logistic regression model that incorporates the main factors by which the two groups differ.

Biases to External Validity

In addition to being internally valid (or adjusted for any biases to internal validity), each study must be designed to estimate the parameter of interest. If the circumstances of a study differ from the circumstances of interest in a way that affects the outcomes of interest, the assessor must again decide whether to include the study in the assessment, and, if so, how to adjust it. The most common threats to external validity are differences in the people who receive the technology (called population bias), differences in the technology they will receive (called treatment intensity), and differences in length of follow-up.

Population Bias. When applied to external validity, a population bias is similar to a selection bias. It occurs when there are systematic differences in the individuals involved in the study compared with the individuals of interest, and these differences can alter the measured effect of the technology.

It is important to understand that not all differences between a study population and the population of interest will alter the estimated effect of the technology. Whether this will occur can depend on which measure of effectiveness is being used. For example, consider an assessment of the effectiveness of screening on breast cancer mortality in average-risk asymptomatic women. Suppose there is a randomized controlled trial involving *high*-risk asymptomatic women. We might be willing to assume that the effectiveness of screening, when measured as the percent reduction in chance of breast cancer death, or as the odds ratio, will be unaffected by the underlying probability that a woman develops breast cancer. This would be the case, e.g., if, for women destined to get breast cancer, screening reduced the probability of death by, say 40%, irrespective of the underlying probability a woman would develop cancer (her incidence rate). On the other hand, if the measure of effect is the absolute difference in probability of dying of breast cancer, then screening a population of women with a relative risk of, say, 2, would have twice the effect as screening a population of women with a relative risk of 1.

Adjustments for population biases affecting external validity are usually performed on the measure of effect. A general model for executing these adjustments is

$$\varepsilon' = \alpha + \beta\varepsilon$$

where ε' is the effect actually estimated for the study, and ε is the effect in the circumstances of interest.

More specific models can be developed to help estimate the parameters α and β, using the logistic regression model as described for selection bias.

Intensity Bias. An intensity bias occurs when there are differences in the technology offered in the study compared with the technology of interest. The dose of

a drug, the frequency of screening examinations, the skill of the practitioners (e.g., efficacy vs. effectiveness), the type of equipment used, and compliance to medications, all might be different in a particular experiment compared with the circumstances of interest.

The approach is to model the outcome in the group that receives the modified technology (θ_1') as a linear combination of the outcome expected in people who receive the technology of interest (θ_1) or no technology at all (θ_0), controlled by an "intensity factor," τ. If the measure of effect is the absolute difference in probability of the outcome, then

$$\theta_1' = \theta_0 + \tau(\theta_1 - \theta_0)$$

If the measure of effect is a relative risk, a suitable model would be

$$\theta_1' = \theta_1^{\tau}\theta_0^{1-\tau}$$

If the measure of effect is the odds ratio, the following model would apply

$$\theta_1' = \left[1 + (\theta_1/(1 - \theta_1)^{\tau}(\theta_0/(1 - \theta_0)^{1-\tau}\right]^{-1}$$

These all have the property that $\tau = 1$ implies $\theta_1' = \theta_1$, and $\tau = 0$ implies $\theta_1' = \theta_0$.

Differences in Length of Follow-Up

Studies frequently follow patients for different lengths of time. If the assessor is willing to assume that the effect measure is independent of length of follow-up (over the range of follow-up periods reported), no adjustment is necessary. If this is not a good assumption, the assessor must model the event rate as a function of time after application of the technology. For example, for many problems it is reasonable to assume that the probability of an event is independent of the length of follow-up. Let p be the number of periods in the circumstances of interest, λ be the arrival rate for repeated events in the circumstances of interest, λ' be the arrival rate for the study. Then $\lambda' = p\lambda$. Other models have been developed (Kaplan Meier, time-to-tumor, etc.).

Uncertainty about Biases

When one is adjusting for biases, it should be recognized that if the assessor is uncertain about the magnitude of any bias, any of the parameters in any of these models can be described as distributions. This uncertainty will automatically be carried through the assessment and reflected in the final distribution for the parameter of interest. If there are dependencies between biases across studies, special implementation techniques are required.

Combinations of Biases

Adjustments can also be nested. For example, if there is dilution in an RCT, the patients who elected not to get the treatment might not be representative of the treatment group as a whole, and they might not have gotten exactly the same intervention as the control group. Analysis of this example would involve three nested adjustments: dilution, selection bias of those who diluted, and technology bias.

For example, let

- α be the proportion of people in the group offered treatment who diluted;
- τ be intensity of the treatment they actually got; and
- β be a patient-selection bias that is additive to the log odds.

Then,

$$\theta'_1 = \alpha \text{Logit}^{-1}(\theta_0 + \tau(\theta_1 - \theta_0)) + (1 - \alpha)\text{Logit}^{-1}(\theta_1 + \beta)$$

where θ_0, θ_1, and θ'_1 are logits of probabilities.

Assigning Weights to Evidence

A method sometimes used to "adjust" pieces of evidence for potential biases is to assign the evidence a "weight," usually between 0 and 1, where 0 causes the evidence to have no influence on the assessment's conclusions, and 1 causes the evidence to be taken at face value. While the Confidence Profile Method contains formulas to implement this type of adjustment, we do not recommend it for two reasons. First, biases cause a piece of evidence to misestimate the parameter of interest, either over- or underestimating it. Weights cannot simulate this basic effect of biases. The use of weights assumes the evidence is still estimating the parameter of interest; the only adjustment it makes is to modify the precision of the estimate. The second reason is that there is no theoretical or evidentiary basis for estimating the appropriate weight to adjust for a specific bias or collection of biases, leaving the choice of a weight arbitrary. The way to avoid these problems is to model the specific effects of each bias, which is the approach taken by the Confidence Profile Method.

INDIRECT EVIDENCE

Much of the available evidence for a technology assessment is indirect. Instead of, or in addition to, directly relating the technology to health or economic outcomes, the evidence might relate the technology to intermediate outcomes. A separate body of evidence must be used to relate the intermediate outcomes to health outcomes. For example, to draw conclusions about the effect of diet on chance of a heart attack, the assessor might need to combine evidence relating diet and serum cholesterol with additional evidence relating serum cholesterol and heart attack rates.

The simplest case assumes that the technology affects the health outcome only through the intermediate outcome. (For example, the only way diet could change heart attack rates is through its effect on serum cholesterol.) The effect of the technology on the health outcome is then the product of two distributions, one relating the technology to the intermediate outcome, and the other relating the intermediate outcome to the health outcome. The Confidence Profile Method also allows for "independent effects" and "transforming effects." The former allows for the possibility that the technology can alter the health outcome independently of any effect on the intermediate outcome. The latter allows for the possibility that the predictive value of the intermediate outcome (as a predictor of the health outcome) can depend on the presence or absence of the technology. For example, the 5-year survival of a stage I cancer detected through screening might be different from the 5-year survival of a stage I cancer detected through signs or symptoms. A specific formula for analyzing indirect evidence involving dichotomous outcomes is described elsewhere (8).

TECHNOLOGY FAMILIES

Often, a collection of experiments will examine several different variations of the technology, in different combinations. For example, one experiment might compare tamoxifen versus a placebo (for treating breast cancer); another might compare tamoxifen versus CMF; another, tamoxifen versus CMF plus radiation; a fourth, CMF plus radiation versus CMF alone; and so forth. The Confidence Profile Method calls these "technology families." Suppose the assessor wants to compare CMF plus radiation versus a placebo. The approach is to use the available evidence to derive probability distributions for the various pairs that have been directly compared. A distribution for the relative effects of other pairs can then be calculated by a series of convolutions. The concept is illustrated by calculating the difference between the test scores of Tom and Bill from knowledge of the differences in scores between Tom and George, and George and Bill.

PROBABILITY DISTRIBUTIONS, CONFIDENCE INTERVALS, AND HYPOTHESIS TESTING

The result of an assessment with the Confidence Profile Method is a joint probability distribution for the parameters of interest. From this, one can calculate the marginal probability distribution for any subset of parameters, and the probability the true value of any particular parameter lies within any specified interval. If the parameters are health or economic outcomes, the distributions can be used directly in value judgments that compare the various effects of a technology, or to compare one technology with another. Because the distributions describe the range of uncertainty about the parameters, they can be used in assessments of (and can be adjusted for) risk aversion. The distributions can also be used as parameters in models to calculate additional outcomes important to a decision.

The probability distributions cannot be used to calculate confidence intervals in the classical sense, although they deliver what in many contexts is more useful information. A 95% confidence interval is the interval such that, if an experiment were repeated an extremely large number of times, the confidence intervals would contain the true parameter 95% of the time. Strictly speaking, this does not mean there is a 95% chance the true value of the parameter lies within these limits for a particular study. Calculation of that type of information requires a probability distribution, as provided by the Bayesian approach. Confidence limits are often mistakingly used for the latter purpose.

Probability distributions also do not test hypotheses or calculate statistical significance, in the classical sense. These techniques ask that the investigator specify in advance a hypothesis about the magnitude of the parameter to be estimated, and then calculate the probabilities that, if that hypothesis were true, the results of the experiment or more extreme results could be expected to occur by chance. If that probability is sufficiently low (traditionally 5% or lower), the results are said to be statistically significant. There are no theoretical limitations on the investigators' choice of the hypothesis to be tested, although the tradition is to use the "null hypothesis" of no effect. A statement about statistical significance, by itself, says little about the actual value of the parameter.

The probability distribution uses a different approach to address the question of whether a technology has an effect. Most important, it can be used to calculate the probability that the effect of the technology is within any range the assessor specifies. For example, from a probability distribution an assessor can calculate the probability

that a treatment decreases the chance of 5-year death by, say, more than 50%, or by 10–20%, and so forth. Because a probability distribution can be used to calculate the probability the actual effect lies in *any* interval, it can calculate the probability the effect is zero or "worse." When a noninformative prior is used, this probability is virtually identical to the classical p value of an experiment.

IMPLEMENTATION

There are two basic approaches to performing an assessment with the Confidence Profile Method: stepwise and integrated. The stepwise approach proceeds by defining chains, identifying links for each chain, identifying the evidence that applies to each link, deriving likelihood functions for each piece of evidence (adjusted for biases as needed), combining evidence for each link, calculating across links for each chain, and combining chains. The process is described elsewhere by Eddy (8). This approach corresponds to our usual way of thinking about pieces of evidence one at a time, and it is intuitively appealing. It also enables the assessor to appreciate the results of each piece of evidence separately. The example at the beginning of this article illustrates the stepwise approach. Software for the stepwise approach is forthcoming.

The stepwise approach has several important limitations. Most important, it is applicable only when there is a single parameter of interest. It also requires that each piece of evidence be independent. While these conditions hold for a large proportion of assessment problems, there are important exceptions. For example, the same bias can affect several studies, or there might be indirect evidence requiring the same evidence to be used as the second link in two chains of evidence. In such cases, the approach is to solve the system of equations that relate all the basic parameters, functional parameters, evidence and prior distributions to obtain a joint probability distribution for all the parameters in the model.

Several methods can be used to solve the model. The solution methods are (a) maximum likelihood estimation (11), (b) exact solution (applicable only for special cases), (c) approximate solution by moments (generally applicable only for a single parameter of interest), and (d) Monte Carlo methods (11). It is also possible to derive a maximum likelihood solution for the parameters of the model, without the use of prior distributions.

In addition to incorporating dependencies between pieces of evidence, this approach also minimizes the requirements on the assessor to determine the appropriate steps required for an assessment, and it might be preferred for complicated assessment problems even when dependencies do not exist. It also generates a wide variety of additional information, such as the correlations between any two parameters.

APPLICATIONS

The Confidence Profile Method has been applied to analyze a variety of assessment problems, including the use of tissue-type plasminogen activator to treat acute myocardial infarction (6;8) and breast cancer screening (7).

SUMMARY

The Bayesian approach as embodied in the Confidence Profile Method has several features. The most notable is that it delivers a probability distribution that can be used directly in decisions or in models to calculate other outcomes. Second, it can handle a variety of experiments, outcomes, and effect measures. For example, it is possible

to combine evidence from a prospective RCT, a retrospective matched case-control study, and a cross-sectional study. Third, it can be used to transform effect measures. For example, the results of an experiment reported as a difference in probabilities can be transformed into an odds ratio or relative risk. Fourth, it allows for adjustment of experiments to incorporate any convictions the assessor might have about factors that affect a variety of biases to internal or external validity. Fifth, uncertainty about any parameter (e.g., an adjustment for a bias that must be estimated subjectively) can be described by a probability distribution, will be automatically combined with other sources of uncertainty according to the axioms of probability, and will be encoded in the final joint distributions for the parameters of interest. There is no need for a sensitivity analysis of these factors. Sixth, it can be used to process indirect evidence.

The Confidence Profile Method differs from classical methods of meta-analysis in several ways. Most important, it enables the assessor to build a model that simultaneously incorporates a variety of parameters. Evidence pertaining to any or all of the parameters can be incorporated. The assessor can define new parameters that are functions of other parameters and can derive distributions for new parameters based on evidence about the original parameters. In this way, conclusions can be drawn about parameters about which there is no direct evidence. In contrast, classical meta-analysis techniques are currently designed to estimate a single parameter for which there is direct evidence. Second, by including prior distributions for the parameters, the method provides a formal technique for incorporating targeted subjective judgments (e.g., "clinical judgment") in an assessment. Third, the method includes a variety of models for adjusting evidence for biases. With rare exceptions, classical techniques take all pieces of evidence at face value. Finally, the Confidence Profile Method is an integrated system permitting calculation of a joint probability distribution function for all the parameters in the model. From this can be calculated marginal distributions for any subset of parameters, and a covariance matrix. The latter is useful for determining where additional research would be most helpful.

NOTES

[1] The term "health outcome" is used to describe an outcome of a disease or injury that affects a person's length or quality of life. Examples are life and death, pain, anxiety, disability, and disfigurement.

[2] Intermediate outcomes are biological changes caused by the illness or treatment that cannot be directly experienced by the patient. Examples are serum cholesterol, intraocular pressure, diameter of an artery, degree of tissue perfusion shown by radioactive dye uptake, and EKG pattern.

[3] The word "experiment" will be used in a general sense to describe any empirical evidence gathered with an explicit design. It will be used interchangeably with "study" and "piece of evidence."

[4] This assumes the results of the two experiments are conditionally independent. If not, the likelihood function for the second experiment must be conditional on the results of the first experiment (i.e., $L(X_2|\theta,X)$.

[5] The inverse of the precision matrix is the classical covariance matrix of the estimates of β_i.

REFERENCES

1. Basu, D. On the elimination of nuisance parameters. *Journal of the American Statistical Association*, 1977, 72, 355-66.
2. Berger, J. O. *Statistical decision theory and Bayesian analysis.* New York: Springer-Verlag, 1984.
3. Berger, J., & Wolpert, R. *The likelihood principle*, 2nd edition. Hayward, CA: Mathematical Statistics, 1988.

4. Breslow, N. E., & Day, N. E. *Statistical methods in cancer research. Vol. 1. The analysis of case-control studies.* Lyon, France: IARC, 1980.
5. DeGroot, M. H. *Optimal statistical decisions.* New York: McGraw-Hill, 1970.
6. Eddy, D. M. The use of confidence profiles to assess tissue-type plasminogen activator. In G. S. Wagner & R. Califf (eds.), *Acute coronary care 1987.* New York: Martinus Nihjoff Publishing Co., 1986.
7. Eddy, D. M., Hasselblad, V., McGivney, W., & Hendee, W. The value of mammography screening in women under age 50. *Journal of the American Medical Association*, 1988, 259, 1512–19.
8. Eddy, D. M. The Confidence Profile Method: A Bayesian method for assessing health technologies. *Operations Research*, 1989, 37, 210–28.
9. Jeffreys, H. *Theory of probability.* London: Oxford University Press, 1961.
10. Laird, N. M., & Mosteller, F. Methods for interpreting and combining experimental results. *International Journal of Technology Assessment in Health Care*, 1990, 6, 5–30.
11. Shachter, R. D., Eddy, D. M., & Hasselblad, V. An influence diagram approach to medical technology assessment. In R. M. Oliver & J. Q. Smith (eds.), *Influence diagrams and belief nets.* New York: Wiley, 1990.
12. Stead, A. G., Hasselblad, V., Creason, J. P., & Claxton, L. Modeling the Ames test. *Mutation Research*, 1981, 85, 13–27.